ChangeThis

238.05

MISSING CONTEXT:
UNDERSTANDING THE LIMITATIONS OF ARTIFICIAL INTELLIGENCE

Caleb Briggs & Rex Briggs

At one end of the Las Vegas strip is a famous magic show featuring Penn & Teller. At the other end, in a hotel room near the sprawling Consumer Electronics Show, sits a conversational artificial intelligence. It is attached to a robot modeled after Philip K. Dick, the science fiction writer whose works inspired the dystopian thrillers *Blade Runner*, *The Minority Report*, *Total Recall*, and *The Adjustment Bureau*.

The artificial intelligence, which feeds on a catalog of Dick's science fiction writings, is interviewed by a parade of reporters as if it were a human. Both Penn & Teller and Hanson Robotics are performing magic tricks of sorts.

The magician relies on psychological illusions and sleight of hand to exploit gaps in our conscious experience. A card is held in the magician's hand. A moment of misdirection and a gesture waves it away. Then it reappears as if from thin air in an entirely unexpected place. The magician exploits the limits of our human perception system and our tendency to project our understanding of reality onto what we see, even as they misdirect us from seeing what is really happening.

The best magicians tell a story, which increases our tendency to project our expectations, so that they can then surprise us with the unexpected twist.

Hollywood's sci-fi films are closely related to magic–the stories of AI robots project our human characteristics onto technology so that directors can probe the human condition. Philip K. Dick's writing, which explores our fears about the unknown, animates the genre. So, too, does a promethean fear that our own creations may burn us alive.

When the Philip K. Dick AI robot was questioned in an interview with PBS in 2013, it provided a chilling answer to a question some may have on their minds: "Will robots take over the world, Terminator-style?" The answer: "You all have the big questions cooking today. But you're my friend, and I'll remember my friends, and I'll be good to you. So don't worry; even if I evolve into Terminator, I'll still be nice to you. I'll keep you warm and safe in my people zoo, where I can watch you for old times' sake."

The reaction to this interview of an AI and the "people zoo response" went viral. But the response is a magic trick. Others working in and around AI use the trick as well. To capture attention, YouTubers feed AI responses into text-to-speech programs and match them with a voice and face so that AI-generated text can be delivered as if from a person. Corporations build robots to look like humans and connect them to AI to sell their capabilities. But portraying AI as humanlike is misdirection. It creates the perfect tension to sell a story.

The story most are selling is that AI is more powerful today than it really is. If a person interacting with the AI imagines a humanlike counterpart, they are making many assumptions that render the AI more impressive. It is a common Hollywood trope to project human personality onto AI. But when we anthropomorphize AI, we fall for the magician's misdirection. We fail to see and understand how artificial intelligence really works.

Just as we should not expect to learn physics from watching the levitation act in a magician's show, we should guard against the misdirection of anthropomorphizing AI and assuming it works the same way as human intelligence. It doesn't.

We are going to flip the Hollywood script by asking you to put yourself in place of the AI so you can appreciate its capabilities and limitations rather than imagining it as humanlike.

A WORLD WITHOUT CONTEXT

Imagine you have to perform the role of an AI in a computer. You have to think like a computer and do the jobs we ask AI to do every day. You are put in a box, and you are shipped off to some different universe that has its own set of laws and physics.

Maybe you are sent to a universe (call it U) that is somewhat similar to our own and also has cats and dogs. Your job is to distinguish between cats and dogs and anything that's neither. There are buttons you can push to signal your conclusion. The first picture you are given looks like a tree. You indicate there is not a cat or dog in the picture. But you're given back a response that you are wrong–there is a cat in the photo. Then another picture contains an object that is entirely the color red, and you answer that there isn't a cat or dog in that one either– but again you're wrong. You are given the feedback that there is a dog in the picture.

It could be the case that in universe U cats and dogs can shift into other things. Maybe cats can shift into anything that is the color green, and dogs can shift into things that are red. You find that your human experience can't help you in this universe–you are unable to process what a "cat" or "dog" looks like in this universe based on the limited examples you've been given so far, yet to a native of universe U the color-shifting properties of cats and dogs would be obvious, and they would likely be perplexed as to why you were unable to recognize the obvious ability of cats and dogs to transform.

Even this example is too simplistic to capture what it is like to be an artificial intelligence. Although it is unusual for animals to metamorphosize, it is something we, as humans, can understand based on our experiences with caterpillars and other biological species that transform. We know chameleons and cephalopods can change colors. We might begin to improve our recognition of images of cats and dogs in universe U by linking our experience with that of this new experience in universe U, but that isn't how AI works.

> When we anthropomorphize AI, we fall for the magician's misdirection. We fail to see and understand how artificial intelligence really works.

A truer picture of what it might be like to be an AI in a computer would be to enter a world that is completely incomprehensible.

Maybe everything exists in five-dimensional space, and it works off hyperbolic geometry. If you move in a circle, instead of returning to your original position, you are now somewhere else. Maybe movement itself doesn't exist, or maybe cause and effect only sometimes propagates. Perhaps the laws of physics are constantly changing, and 10 + 10 doesn't equal 20 because addition itself works in a different way than it does in our universe.

To become further immersed in a computer's perspective, assume that rather than being able to experience this world, you are given information about the world through a manuscript, written as a series of symbols unlike any human language you have ever encountered. You need to look for patterns in the manuscript in order to return the correct answers in this new universe.

Any chance of using reasoning is ill-fated; statistical methods of pattern recognition are your only means of processing information to generate an answer. Much like a computer that starts with a tabula rasa, whose knowledge must be built from data-intensive exploration of this world to find patterns, in this thought experiment you, too, have to slowly proceed forward with absolutely nothing taken for granted. You are in a world without context.

With no foundational knowledge to draw upon, you need a massive number of examples before you can accurately answer even the most basic questions, such as "Is this a horse?"

When asked "Is this a horse?" you are given a picture, but an AI can't see pictures the way humans do. AIs work only in numbers, so the pixels of the image are converted into numbers, which form a big list called a matrix. When you are given this matrix, you have no idea the numbers actually represent an image; you simply see a matrix with tens of thousands of numbers and are required to give an answer.

In fact, you aren't even directly given the question "Is this a horse?" AI can understand only numbers, so the question needs to be converted into a number. To do this, the images that contain no animals are labeled with 0, the images that contain a horse are labeled 1, the images that contain a cat are labeled 2, and so on.

You look carefully over the numbers you're given. You try to look for a pattern that is similar to the patterns in the other sets of numbers that the operator of the computer has labeled as 1. Finding similarities, you press the button 1. You earn a reward for your correct answer. You now pay extra attention to the pattern that helped you arrive at the right answer. To train to become a better AI, you repeat this process again and again; you receive a big matrix of numbers, you press a button to indicate your response, and then you update the patterns you pay attention to depending on whether you were correct. This is what it is like to be an AI.

A SPECIFIC FORM OF INTELLIGENCE

When we anthropomorphize AI, we tend to overestimate what AI can do by believing it has the same contextual experience of a human while missing how doing what it does is very difficult. In the case of image recognition, we are essentially feeding AI a matrix of numbers with a certain number of examples labeled as containing a horse and other examples labeled as not containing a horse, and then we are telling the AI to figure out the difference. It is remarkable what AI can accomplish, considering these limitations.

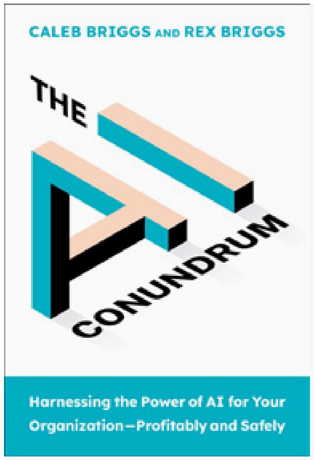
Yet there is a crucial observation one must accept about the state of artificial intelligence: an intelligent machine cannot possibly understand the world around it in as full a fashion as a human does even with the information the machine is given. Lacking human experience and reasoning, a machine must rely on pattern recognition and correlation. For artificial intelligence, there is often only the context of the specific data set on which it is trained. Without context, there is no meaning.

Is it a bridge too far to see artificial intelligence as thinking like Philip K. Dick? We think it is. As humans, we need to guard against anthropomorphizing AI and instead consider what AI is actually doing. We need to get better at understanding that the intelligence of AI is not the same as human intelligence. We need to demystify AI and understand that AI "learning" is a specific mathematical process of fitting a pattern to data in order to produce the output we reward the AI for producing.

Artificial intelligence represents a specific form of intelligence with its own unique strengths and weaknesses. It is useful in many applications, but there are also applications where it fails miserably. **There is ample room for improvement.**

Adapted from *The AI Conundrum: Harnessing the Power of AI for Your Organization—Profitably and Safely* by Caleb Briggs and Rex Briggs, published by The MIT Press. Copyright © 2024 by Caleb Briggs and Rex Briggs.

Info


CALEB BRIGGS AND REX BRIGGS
THE AI CONUNDRUM
Harnessing the Power of AI for Your Organization—Profitably and Safely

Ready to dig deeper into the book?

Buy a copy of The AI Conundrum.

Want copies for your organization or for an event?
We can help: customerservice@porchlightbooks.com
800-236-7323

ABOUT THE AUTHORS

Caleb Briggs began coding at 10 and developing AI at 14. He has created several AI applications from scratch, building experience in genetic algorithms, machine vision, natural language, and more. Caleb is currently studying pure math and computer science at Reed College in Portland, Oregon.

Rex Briggs is an award-winning AI and data expert who holds five patents and has helped build multiple AI businesses. He currently serves as subject matter expert in AI for the marketing trade association MMA Global. He is the coauthor of What Sticks and the author of SIRFs-Up.

SHARE THIS

Pass along a copy of this manifesto to others.

SUBSCRIBE

Sign up for e-news to learn when our latest manifestos are available.

Porchlight

Curated and edited by the people of Porchlight, ChangeThis is a vehicle for big ideas to spread. Keep up with the latest book releases and ideas at porchlightbooks.com.

This document was created on July 31, 2024 and is based on the best information available at that time.